Evolution patterns and parameter regimes in edge localized modes on the National Spherical Torus Experiment

Plasma Phys. Control. Fusion **58** (2016) 045003 (16pp) doi:10.1088/0741-3335/58/4/045003

Evolution patterns and parameter regimes in edge localized modes on the National Spherical Torus Experiment

D R Smith[1], R J Fonck[1], G R McKee[1], A Diallo[2], S M Kaye[2], B P LeBlanc[2] and S A Sabbagh[3]

[1] Department of Engineering Physics, University of Wisconsin-Madison, Madison, WI 53706, USA
[2] Princeton Plasma Physics Laboratory, Princeton, NJ 08543, USA
[3] Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA

E-mail: drsmith8@wisc.edu

Received 14 July 2015, revised 9 December 2015
Accepted for publication 21 December 2015
Published 28 January 2016

Abstract

We implement unsupervised machine learning techniques to identify characteristic evolution patterns and associated parameter regimes in edge localized mode (ELM) events observed on the National Spherical Torus Experiment. Multi-channel, localized measurements spanning the pedestal region capture the complex evolution patterns of ELM events on Alfvén timescales. Some ELM events are active for less than 100 μs, but others persist for up to 1 ms. Also, some ELM events exhibit a single dominant perturbation, but others are oscillatory. Clustering calculations with time-series similarity metrics indicate the ELM database contains at least two and possibly three groups of ELMs with similar evolution patterns. The identified ELM groups trigger similar stored energy loss, but the groups occupy distinct parameter regimes for ELM-relevant quantities like plasma current, triangularity, and pedestal height. Notably, the pedestal electron pressure gradient is not an effective parameter for distinguishing the ELM groups, but the ELM groups segregate in terms of electron density gradient and electron temperature gradient. The ELM evolution patterns and corresponding parameter regimes can shape the formulation or validation of nonlinear ELM models. Finally, the techniques and results demonstrate an application of unsupervised machine learning at a data-rich fusion facility.

Keywords: edge localized modes, time series analysis, unsupervised machine learning, national spherical torus experiment (NSTX)

(Some figures may appear in colour only in the online journal)

1. Introduction

The linear peeling-ballooning stability boundary expresses an onset condition for edge localized modes (ELMs) [1], but ELM saturation mechanisms, filament dynamics, and multi-mode interactions require nonlinear models [2–5]. For instance, models for edge harmonic oscillations point to low-n peeling modes destabilized by rotational shear and held in a saturated state through rotational damping from wall drag [6, 7]. Nonlinear magnetohydrodynamic simulations indicate hyper-resistivity is a key factor for realistic radial penetration during the pedestal collapse [4]. Also, nonlinear simulations indicate low-n modes that are subdominant during the linear phase can grow to amplitudes that rival dominant modes due to nonlinear mode coupling [2, 5]. Typical diagnostic tools for ELM observations, like Thomson scattering profiles and D_α filterscopes, do not resolve the Alfvén timescales of ELM events. Furthermore, heuristic ELM classification schemes (Type I, III, etc) based on extrinsic ELM properties, like secular edge emission and inter-ELM period, do not capture the nonlinear dynamics and Alfvén-scale evolution of ELM events [8]. Validation of nonlinear ELM models requires fast measurements on Alfvén timescales, and successful models should reproduce the complex evolution patterns observed during ELM events. Identification of common evolution patterns in ELM events can motivate the formulation or validation of nonlinear ELM models.

In this paper, we investigate Alfvén-scale evolution patterns in ELM events captured by beam emission spectroscopy (BES) measurements [9, 10] on the National Spherical Torus Experiment (NSTX) [11, 12]. We implement unsupervised machine learning algorithms that identify groups of ELMs with similar time evolution characteristics. The ELM database in this investigation most likely includes only Type I ELMs due to the ELM selection criteria. The analysis points to at least two and possibly three ELM groups with distinct evolution patterns, which suggests recurring and distinct variations in nonlinear dynamics or saturation mechanisms. The identified ELM groups exhibit similar stored energy loss, but the groups occupy distinct parameter regimes for plasma current, triangularity, magnetic balance, and pedestal height. The observed evolution patterns and associated parameter regimes can motivate nonlinear ELM models or validation scenarios for ELM simulations. Finally, several scientific fields leverage machine learning techniques to automate scientific discovery tasks like pattern identification, data classification, or relationship quantification in large, complex datasets for which exhaustive visual inspection of data is not feasible. The analysis presented here demonstrates an application of unsupervised machine learning at a data-rich experimental fusion facility.

In the remainder of this paper, section 2 describes the NSTX ELM database and presents example ELMs. Section 3 presents unsupervised machine learning analysis (hierarchical clustering and *k*-means clustering) that identifies ELM groups with similar evolution patterns, and section 4 considers plasma parameter regimes that correlate with the identified ELM groups. Section 5 discusses opportunities to leverage machine learning techniques and large data archives at large experimental fusion facilities. Section 6 provides a summary of results. Finally, digital data for all figures and analysis in this manuscript can be found in [13].

2. ELM event database

We identified 51 ELM events from the NSTX data archive with beam emission spectroscopy (BES) measurements spanning pedestal region and into the core plasma, as shown in figure 1. BES measurements of plasma density are localized with $\Delta x \approx 2$ cm and sample on Alfvén timescales at 2 MHz ($\Delta t = 0.5\ \mu$s, $\tau_A \sim 5\ \mu$s, and $\Delta t/\tau_A \sim 0.1$) [9, 10]. The measurement locations are fixed in space, so the locations in normalized flux can change with plasma shaping and position control. For this reason, we utilized the radial array of BES sightlines in figure 1. The radial array covers the pedestal region in all discharge scenarios. As shown in figure 2, BES measurements capture the fast nonlinear evolution of ELM events in contrast to conventional filterscope measurements or Thomson scattering profiles. As we describe below, the multiple BES signals are condensed into a single, representative time-series with principle component analysis.

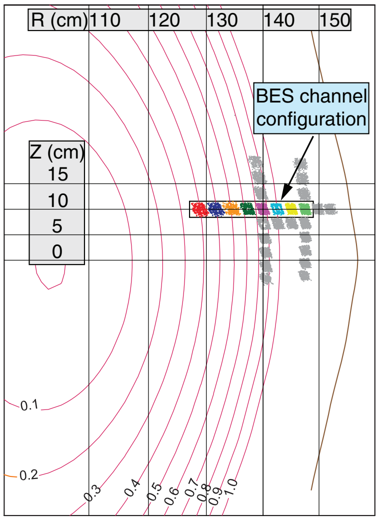

Figure 1. NSTX cross-section showing BES channels in a radial configuration of ELM observations. Contour labels are normalized poloidal flux.

The ELM database was populated with the following objectives: (1) sample ELMs for a variety of machine and wall conditions, (2) identify ELMs that are isolated from other ELMs (ELM periods $\gtrsim$30 ms) and confounding MHD activity, such as Alfvén avalanches, and (3) include only ELMs that exhibit a clear pedestal collapse or stored energy loss. An ELM pedestal collapse may not be observable if the ELM occurred early in the period between Thomson scattering measurements, and the stored energy loss may be erroneously small if the ELM occurred early in the period between magnetic reconstructions. Therefore, we require either a clear pedestal collapse or stored energy loss for ELMs in the database. The constraints likely exclude small, grassy, or Type V ELMs and Type III ELMs with periods $\lesssim$20 ms. In other words, the ELM database is likely populated only by Type I ELMs. To capture diverse ELM phenomena in a variety of machine conditions, the 51 ELM events were drawn from 34 H-mode discharges spanning four months of experimental operations. The ELM events show stored energy losses up to 16%. As shown in table 1, the ELM database spans a large range of plasma current, auxiliary heating power, plasma shape, and magnetic topology.

Measurements on Alfvén timescales inherently capture the nonlinear dynamics and saturation mechanisms of ELM events, and figure 3 shows examples of diverse ELM events in the database. For instance, some ELM events last less than 100 μs, but others persist up to 1 ms. A single perturbation dominates some ELM events, but other events show multiple perturbations. Finally, some events are oscillatory, but others are non-oscillatory. At this point, any structure, organization, or pattern in the ELM database is unknown. Structure in the ELM database

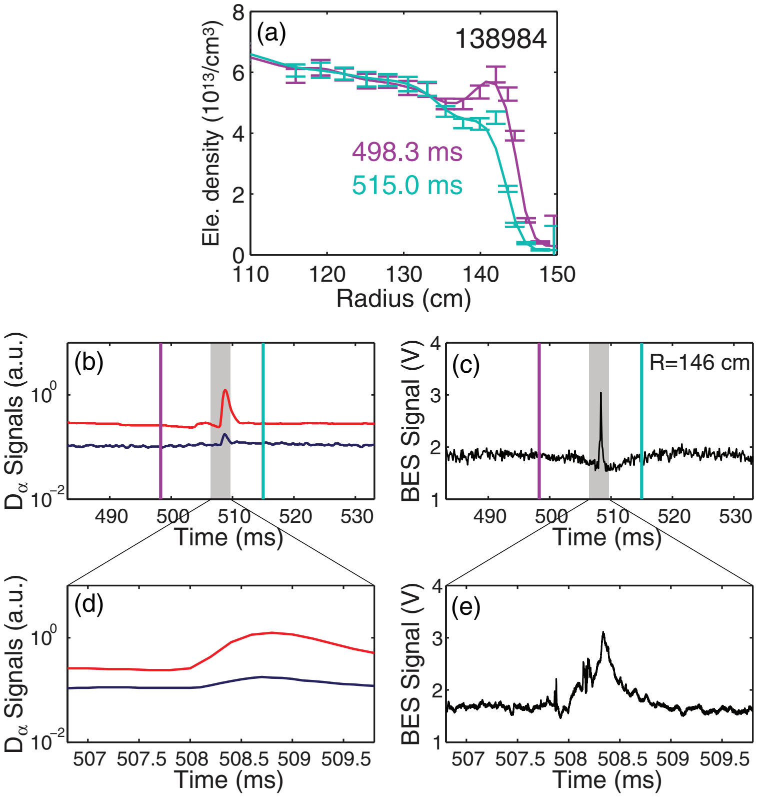

Figure 2. An example ELM event: (a) Thomson scattering profiles of electron density before and after the ELM event; (b) and (d) D_α filterscope signals viewing inner diverter (red) and full diverter (blue); and (c) and (e) BES measurements of the ELM event at $R = 146$ cm. In (b) and (c), vertical lines denote Thomson scattering measurement times.

might point to common nonlinear processes that govern ELM evolution, so our first objective is to identify any structure or pattern in the database. Often visual inspection is adequate to identify structure in data, but visual inspection is not scalable to large or high-dimensional datasets. Unsupervised machine learning algorithms can identify structure, patterns, or organization in unlabeled data with computational speed and scalability. In the next section, we implement unsupervised machine learning algorithms to search for patterns in the ELM database.

3. Cluster analysis of ELM time series data

Unsupervised clustering techniques can identify structure, patterns, or association in datasets. Here, we begin with hierarchical clustering for the ELM dataset described in the previous section, and later we explore *k*-means clustering [14]. Hierarchical clustering links data objects in a multi-level hierarchy according to the degree of similarity, and clustering is a common tool in genomics for linking gene expression and functional pathways [15]. The hierarchical clustering algorithm operates on a distance-like metric that quantifies dissimilarity between data objects, and the time series ELM data we examine requires metrics that quantify dissimilarity between time series [16]. Time series similarity metrics operate on a pair of time series, so we apply principal component analysis (PCA) to reduce the multi-channel BES data (see figure 3) to a single representative time series for each ELM event. PCA, a common technique for dimensionality reduction, is an orthogonal coordinate transformation in which the first coordinate axis (first principal component) is the data projection with maximum variance. Figure 4 shows examples of PCA time series extracted from multi-channel BES data. The first principal component ('PC 1') is representative of evolution patterns in all BES channels, and the analysis below is performed on the PC 1 time series for each ELM event. PC 1 typically captures over 74% of the variance in the signals, and PC 2 typically accounts for less than 17% of the variance.

The time-lag cross-correlation (TLCC) is the first dissimilarity metric we consider. For ELM time series $X(t)$ and $Y(t)$, the time-lag cross-correlation, $\rho(\tau)$, is

$$\rho(\tau) \equiv \frac{1}{N_\tau} \sum_{t=1}^{N_\tau} \frac{(X(t+\tau) - \mu_X)(Y(t) - \mu_Y)}{\sigma_X \sigma_Y} \qquad (1)$$

where τ is the time delay, and N_τ is the time delay record length, σ is the time series standard deviation, and μ is the time series mean value. Figures 5(a)–(d) show ELM events with similar and dissimilar time evolution and time lag cross-correlations. As expected, the pair of similar ELMs exhibit the larger $\max(\rho)$ value. The hierarchical clustering algorithm operates on dissimilarity metrics, so maximum correlation values are converted to dissimilarity values with

Table 1. 10th–90th percentile ranges for plasma parameters in the ELM database.

Parameter	Parameter range
Plasma current, I_p (MA)	0.6–1.2
Toroidal field, B_t (kG)	4.4–5.4
Neutral beam power, P_{nb} (MW)	2.9–4.8
Magnetic balance, d_r^{sep} (cm)	−1.33 to 0.14
Safety factor, q_{95}	6.1–11.2
Stored energy, W_{dia} (kJ)	127–260
Stored energy loss, W_{loss} (%)	0.7–9.9
Elongation, κ	1.9–2.4
Lower triangularity, δ_l	0.46–0.73
Pedestal ele. den. height, n_e^{ped} (10^{13}/cm^3)	4.2–6.0
Pedestal ele. den. width, $\Delta R_{ne}^{\text{ped}}$ (cm)	3.2–10.1
Pedestal ele. temp. height, T_e^{ped} (keV)	0.33–0.67
Pedestal ele. temp. width, $\Delta R_{Te}^{\text{ped}}$ (cm)	5.8–11.3
Z_{eff}	1.5–3.2
Collisionality, ν_{ei} (10^6)	1.3–4.2
Total beta, $\beta = 2\mu_0(2p_e^{\text{ped}})/B^2$	0.16–0.40
Poloidal beta, $\beta_p = 2\mu_0(2p_e^{\text{ped}})/B_p^2$	0.68–3.3
Alfvén time, $\tau_{\text{A}} = R_0/v_{\text{A}}$ (μs)	4.3–5.6

Note: The 10th–90th percentile range captures approximate minimum and maximum values while excluding distortion from extreme outliers. Digital data for the ELM database can be found in [13].

$$D_{\text{TLCC}} \equiv 1 - \max(\rho(\tau)) \tag{2}$$

where D_{TLCC} is the dissimilarity metric for time-lag cross-correlation. The hierarchical clustering algorithm operates on a dissimilarity matrix containing dissimilarity values for all ELM pairs. Figure 5(e) shows the dissimilarity matrix for D_{TLCC} with blue colors indicating ELM pairs with low dissimilarity (high similarity), and the matrix is necessarily symmetric with respect to ELM index.

The hierarchical clustering algorithm iteratively merges the most similar (least dissimilar) data objects into clusters, and the output is a multi-level hierarchy with the most similar objects linked at low levels. As new clusters are created, a linkage formula sets dissimilarity metrics for new clusters relative to other clusters, and complete linkage is the maximum distance between objects in the new cluster and objects in other clusters. For instance, if the clustering algorithm creates a new cluster p, then the complete linkage between new cluster p and an existing cluster q is

$$L_{\text{com}}(p,q) \equiv \max(D(x_i^p, x_j^q)) \tag{3}$$

where x_i^p are the objects in cluster p and x_j^q are the objects in cluster q. Figure 6 shows the results of hierarchical clustering with complete linkage for the ELM dissimilarity matrix from figure 5(e). Dendrograms illustrate the multi-level hierarchy of data objects and clusters, and figure 6(a) shows the dendrogram for the ELM database with complete linkage. Groups of data objects linked with high similarity in the dendrogram are candidate clusters, and figure 6(a) shows three candidate clusters labelled 1 (red), 2 (blue), and 3 (green). Subsequent analysis in this section will demonstrate that clusters 1, 2, and 3 persist across multiple clustering techniques and algorithms. The designation of clusters in dendrograms is somewhat subjective, but good candidate clusters should preferably contain many members with high similarity inside the cluster and low similarity outside the cluster. We feel the identified clusters in figure 6(a) best captures the criteria for good clusters. If less similarity within clusters is tolerable, then cluster 1 can expand to include ELMs 7/40/26/42/47. At even lower similarity tolerance, clusters 1 and 2 merge leaving two dominant clusters. In hierarchical clustering, some data objects may not clearly belong to a cluster, as evident in figure 6(a). The hierarchical clustering algorithm iteratively merges the data objects or clusters with the lowest linkage, so it is feasible that some data objects are isolated with large linkage values to a larger cluster, like ELMs 48 and 49 in figure 6(a). In contrast, in k-means clustering (discussed below), all data objects are assigned to a cluster. For an alternative visualization, the dissimilarity matrix figure 5(e) can be reordered according to the data sequence in the dendrogram, as shown in figure 6(b). In the reordered dissimilarity matrix, candidate clusters are square regions along the diagonal with low dissimilarity values, and the three candidate clusters are designated with colored squares in figure 6(b). Figures 6(c)–(h) shows examples of ELM evolution for the identified clusters. ELMs in cluster 2 are short duration (~30 μs), and cluster 1 ELMs are similarly intense but with longer duration (~400 μs). Finally, ELMs in cluster 3 show elevated signals that persist over 1 ms. In summary, the clustering algorithm, without prior training or user guidance, created a hierarchy of ELM events grouped by the degree of similarity. The modest results in figure 6 are notable because (1) identifying structure, patterns, or association in data is a fundamental activity in scientific discovery; (2) algorithmic pattern recognition is scalable and readily automated; and (3) experimental fusion facilities generate large volumes of data. The remainder of this section shows that the hierarchical clustering results in figure 6 hold for other linkage formulas and dissimilarity metrics, and the results are consistent with k-means cluster analysis. The consistency of results from multiple algorithms and metrics adds credibility to the identified ELM evolution patterns. The next section (section 4) explores parameter regimes that correspond to the ELM clusters identified in figure 6.

Figure 6 showed hierarchical clustering results for complete linkage, and figure 7 shows similar results using an average linkage algorithm in which dissimilarity metrics for new clusters are the average distance between objects in the new cluster and other clusters. Specifically, the average linkage between a new cluster p and an existing cluster q is

$$L_{\text{avg}}(p,q) \equiv \frac{1}{n_p n_q} \sum_{i=1}^{n_p} \sum_{j=1}^{n_q} D(x_i^p, x_j^q) \tag{4}$$

where x_i^p is the ith object in cluster p with n_p objects and x_j^q is the jth object in cluster q with n_q objects. Average linkage values in figure 7 are less than complete linkage values in figure 6 because the average distance between data objects in two clusters is necessarily less than the maximum distance between the data objects. The clusters and colors in figure 7

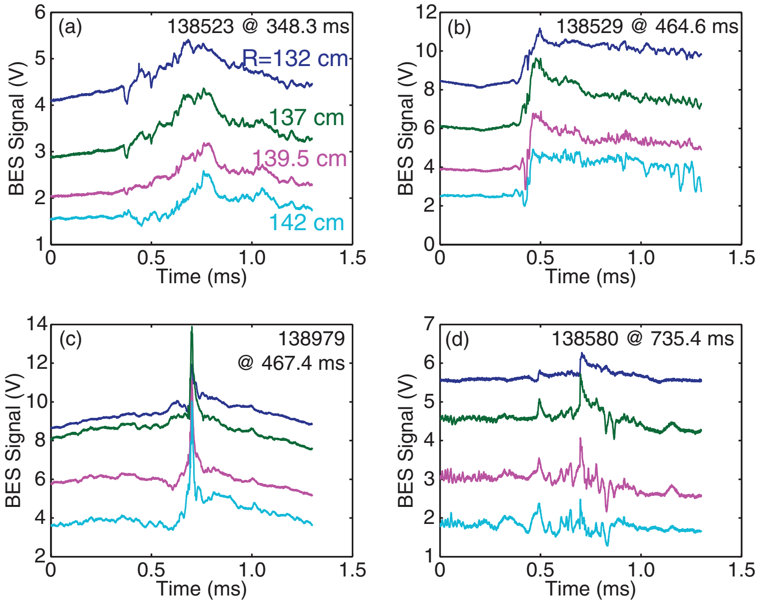

Figure 3. (a)–(d) Examples of nonlinear ELM evolution captured by multi-point BES measurements with high time resolution. Shown are four of eight BES signals from the radial array highlighted in figure 1, and the colors correspond to measurement locations in figure 1. The data segments are chosen to capture the fast evolution of the entire ELM burst, but $t = 0$ selection is not systematic due to the lack of a universal marker for syncing across ELM events. The time-series analysis presented in section 3 is insensitive of $t = 0$ selection. BES digital data for the ELM database are available in [13].

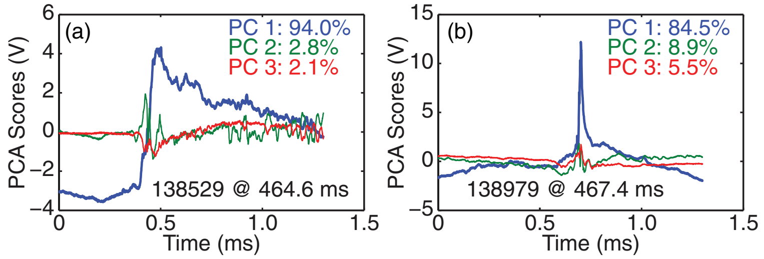

Figure 4. Top three principal components (PC) and variance percentages for BES signals in (a) figure 3(b) and (b) figure 3(c). The largest principal component ('PC 1') is representative of dominant evolution patterns in radial BES signals.

are preserved from figure 6, and the clusters' memberships are largely preserved.

Figure 6 showed results for the time-lag cross-correlation dissimilarity metric, but now we repeat the analysis with other dissimilarity metrics (time-lag Euclidean distance (TLED), dynamic time warping, and the time-lag cross-correlation) to assess validate the results. The TLED is the minimum root-mean-square distance as a function of time delay,

$$D_{\mathrm{TLED}} \equiv \min_{\tau}\left(\sqrt{\frac{1}{N_\tau}\sum_{t=1}^{N_\tau}(X(t+\tau)-Y(t))^2}\right). \tag{5}$$

Figures 8(a)–(b) show examples of TLED for similar ELMs (blue) and dissimilar ELMs (red). In figure 8(b), the minimum TLED is lower for the blue ELMs with similar time evolution. Dynamic time warping (DTW) is similar to the TLED, but DTW allows timebase distortion to minimize a dissimilarity cost function. For ELM time series $X(i)$ and $Y(j)$ with $i, j \leqslant T$, the DTW distance is calculated from the recursive algorithm

$$D_{i,j} = f_{i,j} + \min(D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}) \tag{6}$$

with cost function $f_{i,j} = |X(i) - Y(j)|$ and initial conditions $D_{i,0} = D_{0,j} = \infty$. The DTW distance metric is $D_{DTW} \equiv D_{T,T}$. Figure 8(c) shows a pair of similar ELMs, and figure 8(d) shows the corresponding DTW calculation. $D_{T,\,T}$ is upper-right point in figure 8(d), and we see $\log(D_{T,T}) \approx -4$ for the similar ELMs. For comparison, figure 8(e) shows dissimilar ELMs, and figure 8(f) indicates $\log(D_{T,T}) \approx -2$.

The final dissimilarity metric we consider is the time-lag cross-correlation of wavelet-transformed signals. We perform a multilevel discrete wavelet decomposition [17] to capture both short, rapid changes and slower trends in the signals. As the iterative wavelet decomposition advances, coarser signal trends are extracted. We apply the Daubechies db4 wavelet, a high-pass finite-impulse-response filter with four vanishing

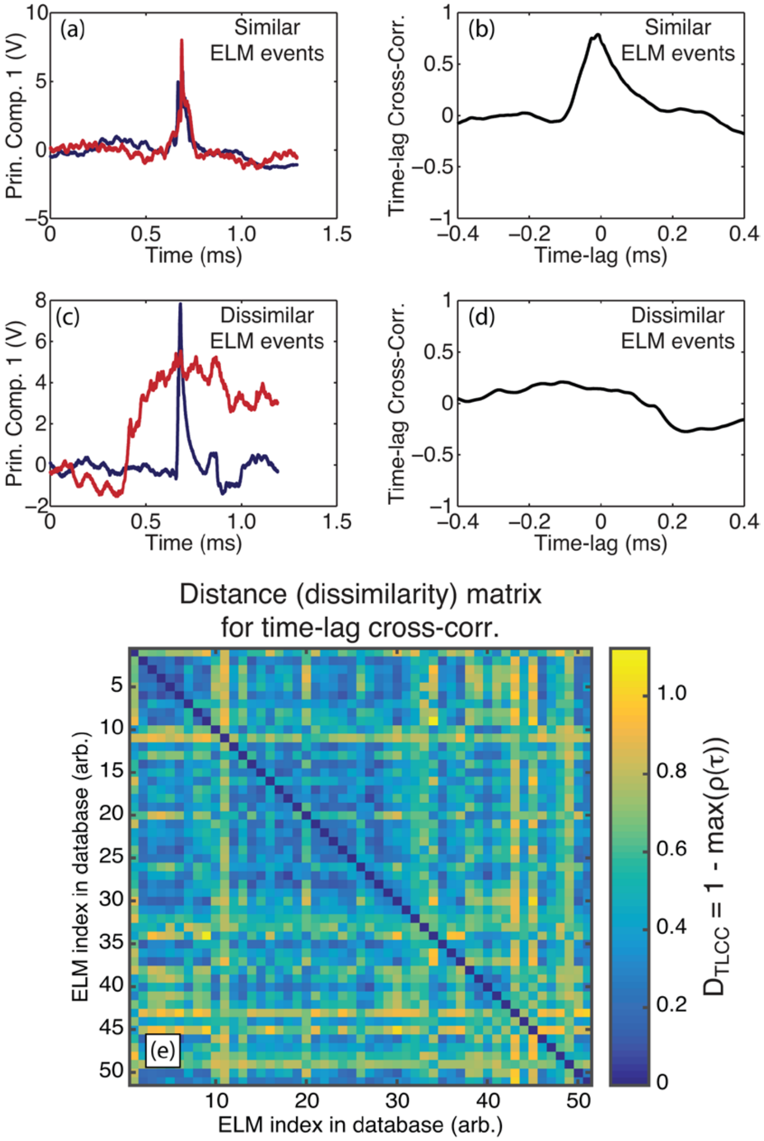

Figure 5. ((a), (b)) Similar ELMs exhibit a large maximum time-lag cross-correlation, and ((c), (d)) dissimilar ELMs exhibit a small maximum time-lag cross-correlation. (e) The maximum time-lag cross-correlation for all ELM pairs can be visualized as a symmetric dissimilarity matrix.

moments. Convolutions with db4 are principally sensitive to signal segments that grow faster than t^4, including exponential growth. At each level in the decomposition, the signal is convolved with the high-pass db4 ('detail' output) and the low-pass quadrature mirror of db4 ('approximation' output). The detail and approximation outputs are decimated by two, and the approximation output is the input signal for the next level in the wavelet decomposition. For signal s, the iterative wavelet decomposition with db4 is

$$a_0 \equiv s \tag{7}$$

$$d_i \equiv \mathrm{dec}(\mathrm{conv}(a_{i-1}, \mathrm{db4})) \qquad (i \geqslant 1) \tag{8}$$

$$a_i \equiv \mathrm{dec}(\mathrm{conv}(a_{i-1}, \mathrm{quad}(\mathrm{db4}))) \qquad (i \geqslant 1) \tag{9}$$

where d_i is the ith detail, a_i is the ith approximation, dec denotes decimation by two, conv denotes convolution, and quad denotes quadrature mirror transformation. The approximations are a smoothed versions of the original signal, and the details capture signal deviations faster than t^4. To construct a dissimilarity metric between two ELM events, we calculate time-lag cross-correlations between detail and approximation signals for ELM , and then we convert the correlations into dissimilarity values like equation (1). We found that the level 2 detail signal and the level 5 approximation signal captured ELM groups generally consistent with previous dissimilarity metrics. For example, figure 8(g) shows similar (blue) and dissimilar (red) ELM events, and figure 8(h) shows time-lag cross-correlations between the similar and dissimilar ELMs events for the level 2 detail signals and level 5 approximation

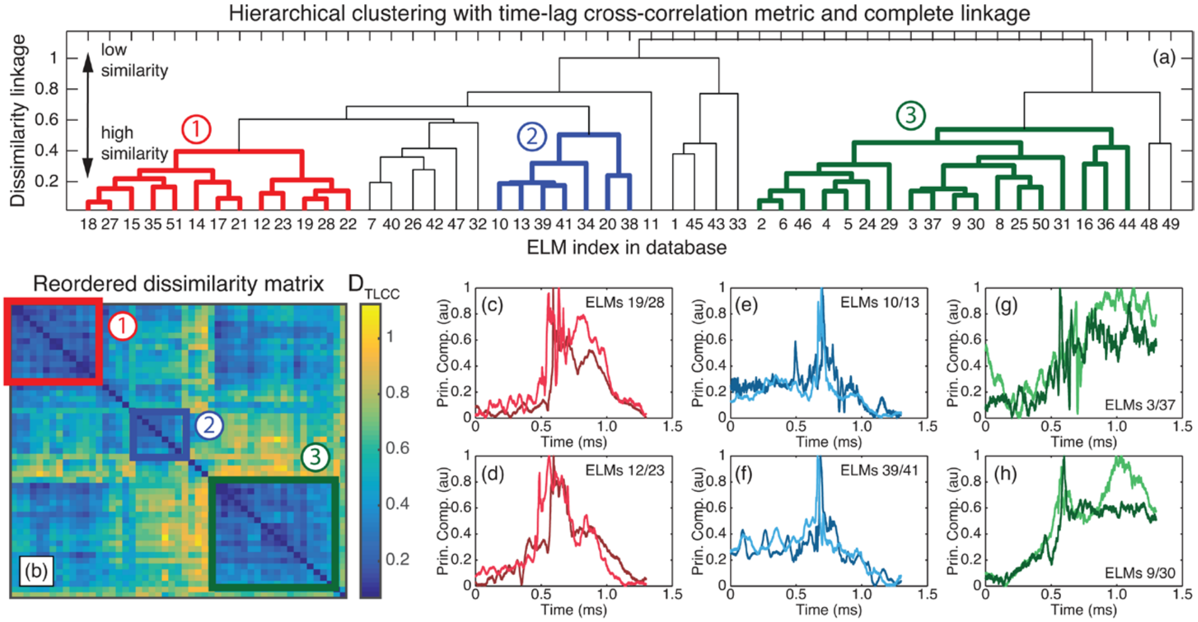

Figure 6. (a) Dendrogram showing hierarchical clustering in the ELM database with the time-lag cross-correlation dissimilarity metric and complete linkage, and (b) the dissimilarity matrix reordered for the ELM sequence in the dendrogram. Clusters 1 (red), 2 (blue), and 3 (green) denote groups of ELMs with similar evolution characteristics. (c)–(h) Examples of similar ELMs from the clusters. Digital data for the ELM database can be found in [13].

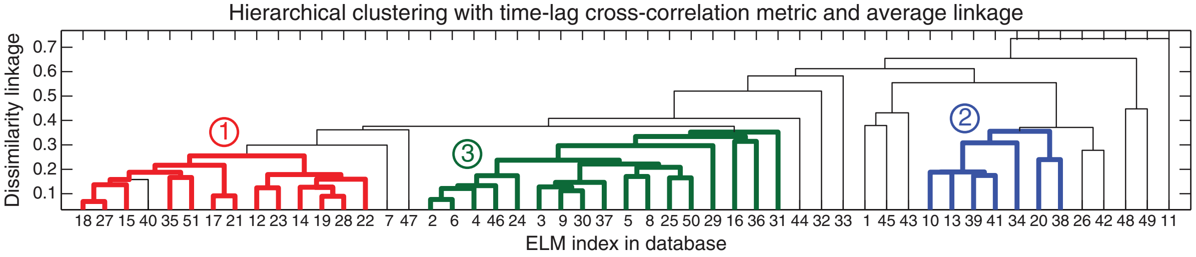

Figure 7. Dendrogram showing hierarchical clustering in the ELM database with the time-lag cross-correlation dissimilarity metric and average linkage.

signals. The similar ELM events show larger peak correlations in both detail and approximation signals.

With several dissimilarity metrics at hand, we can now compare hierarchical clustering results for different metrics. Figure 9 shows hierarchical clustering results with average linkage for (b)–(e) four dissimilarity metrics and (a) the geometric mean of the four dissimilarity metrics. Like figure 6, three clusters of similar ELM events emerge for the geometric mean clustering. The identified clusters are linked at low linkage values and are largely preserved for different dissimilarity metrics. The metrics for (b) time-lag cross-correlation, (c) TLED, and (d) dynamic time warping produce similar results that are largely consistent with (a) geometric mean clustering. Cluster results for (e) wavelet decomposition, however, are the least consistent with other metrics and the geometric mean clustering. Hierarchical clustering with the wavelet metric split cluster 1 and barely catches any hint of cluster 2. Consistent clustering results from three metrics in figures 9(b)–(d) is encouraging despite low consistency from the wavelet metric in (e). Next we examine *k*-means clustering, and we find that *k*-means clustering results are consistent with figures 9(a)–(d) which further boosts confidence that the evolution patterns represent meaningful variations in ELM dynamics.

Like hierarchical clustering, *k*-means clustering is an unsupervised learning algorithm that identifies structure, patterns, or association in data. A key difference between the clustering techniques is that *k*-means clustering assigns all data objects to a cluster. In contrast, hierarchical clustering can yield outliers that do not belong to a recognizable cluster. Finally, hierarchical clustering operates on relative distance metrics, but *k*-means clustering requires absolute coordinates. Here, we designate a set of benchmark ELMs, and dissimilarity metrics for the benchmark ELMs function as absolute coordinates in the *k*-means algorithm. Figure 10(a) shows the geometric mean of the four dissimilarity metrics (correlation, Euclidean, DTW, and wavelet) for six benchmark ELMs.

The number of clusters is a specified parameter in the *k*-means algorithm, and the optimum cluster number is found through

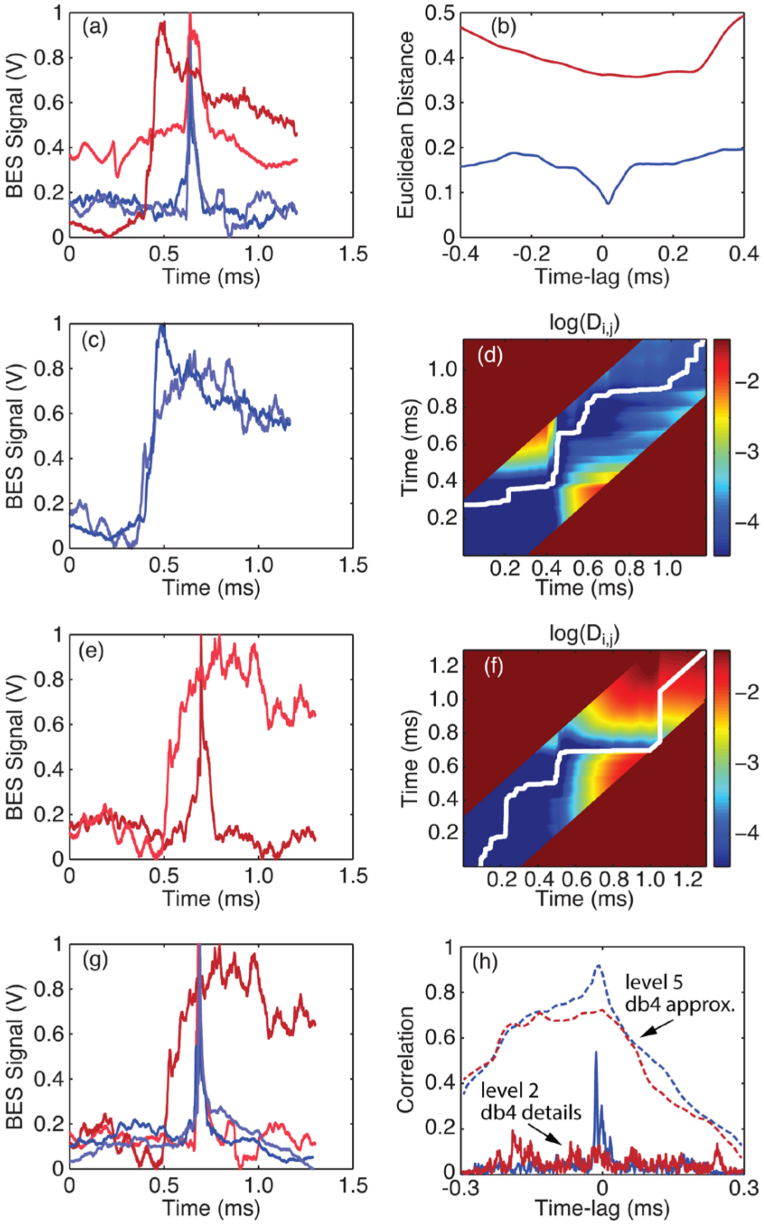

Figure 8. (a) Similar (blue) and dissimilar (red) ELMs, and (b) the associated Euclidean distance similarity metrics. (c) Similar ELMs and (d) the DTW calculation. (e) Dissimilar ELMs and (f) the DTW calculation. (g) Similar (blue) and dissimilar (red) ELMs, and (h) the time-lag cross-correlations for the level 2 db4 detail signals (solid) and level 5 db4 approximation signals (dashed). Digital data for the ELM database is available in [13].

trial-and-error. The 'silhouette' value for each data object measures similarity to objects in its own cluster relative to objects in other clusters. For data object i, the silhouette value is

$$s_i \equiv \frac{b_i - a_i}{\max(a_i, b_i)} \tag{10}$$

where a_i is the average dissimilarity with other objects in the same cluster and b_i is the minimum dissimilarity with other clusters. Large s_i values indicate the data object is appropriately clustered. The optimum cluster number for the k-means algorithm is the cluster number J that maximizes

$$S_J \equiv \text{mean}(s_i). \tag{11}$$

The S_J values in table 2 indicate that four clusters are optimum. Visualizing four clusters of six-dimensional data is difficult, but plotting results in a subspace of two or three principle components aids visualization. Principal component analysis in figure 10(b) indicates that the first three principal components capture nearly 95% of variation in the six-dimensional data in figure 10(a) for the benchmark ELMs. Figure 11 illustrates the optimum four clusters plotted in terms of principal components. Clusters 1,

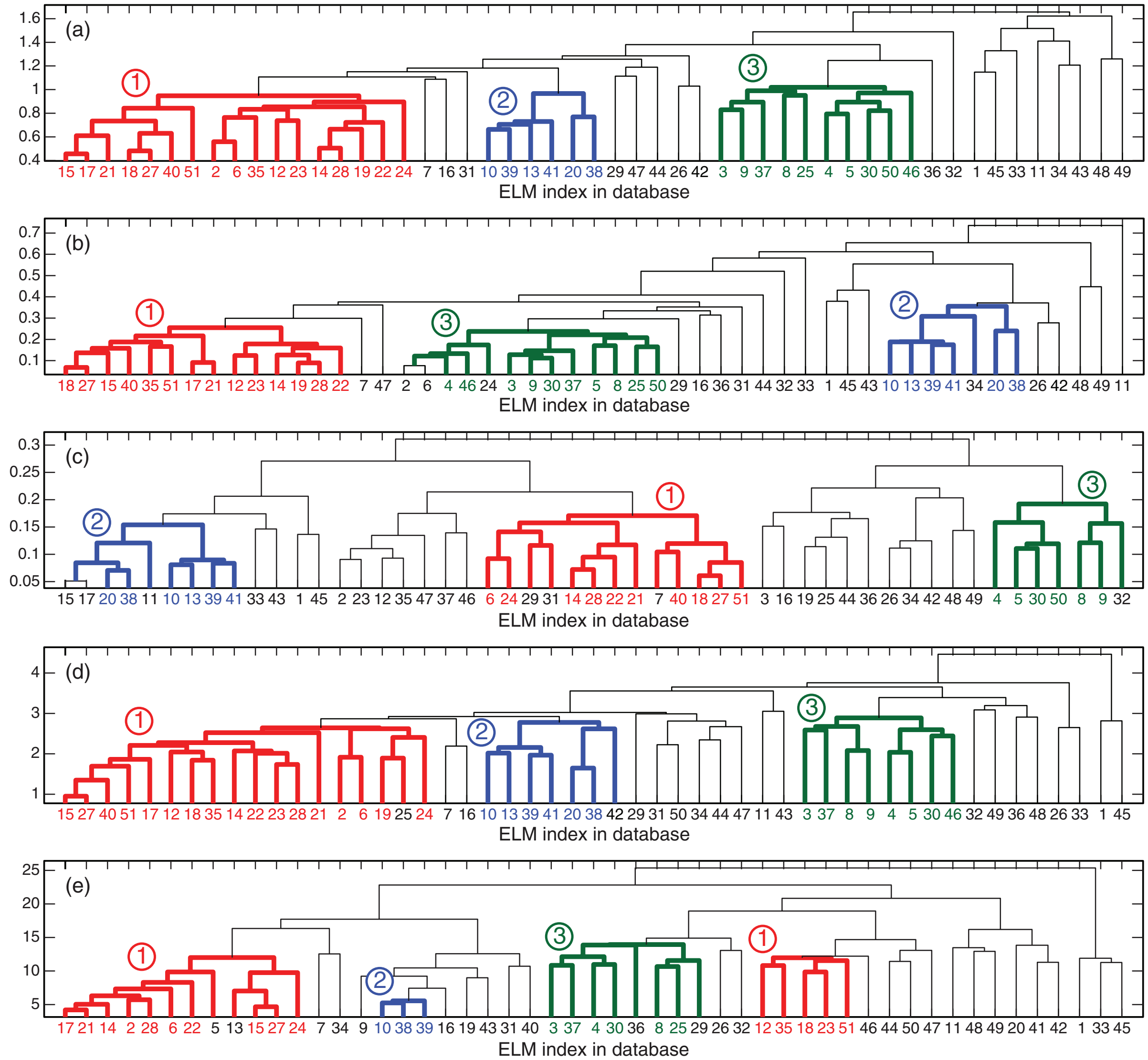

Figure 9. (a) Hierarchical clustering using the geometric mean of four dissimilarity metrics: (b) time-lag cross-correlation, (c) TLED, (d) dynamic time warping, and (e) cross-correlation of wavelet-transformed signals.

2, and 3 in figure 11 correspond to the same clusters in figure 9, but the cyan cluster in figure 11 has no corresponding low-linkage cluster in figure 9 For this reason,w e are reluctant to attached cluster '4' designation to the cyan cluster in figure 11. Later in this section, we will find that clusters 1–3 in figure 11 map to corresponding clusters with low linkage values from hierarchical clustering (figures 6, 7, and 9), plus the significance of the cyan cluster in figure 11 will be explained.

To validate the cluster results in figure 11, we repeat the calculation with additional benchmark ELMs. Table 3 lists the optimal cluster number and S_J values for 6, 9, 11, and 14 of benchmark ELMs, and the results indicate four clusters are optimal for all benchmark scenarios. In all benchmark scenarios, principal components 1–3 captured at least 87% of variation in the benchmark data, so plotting results in principal component space is still effective for visualization.

Table 2. S_J values for the six benchmark ELMs in figure 10.

Number of clusters, J	2	3	**4**	5	6	7
S_J	0.49	0.51	**0.52**	0.48	0.46	0.45

The four clusters for each benchmark scenario are illustrated in figure 12 in terms of principal components, and the clusters' memberships are nearly identical across the benchmark scenarios. Therefore, *k*-means clustering calculations yield four clusters of ELM events, and the clusters' memberships are robust for different sets of benchmark ELMs.

Now we tie together results from *k*-means clustering and hierarchical clustering. The *k*-means clustering results in figures 11 and 12 indicate four clusters are optimal, but the hierarchical clustering results in figures 6, 7 and 9 point to three clusters of ELMs with similar evolution. The apparent discrepancy is resolved by mapping *k*-means results to the

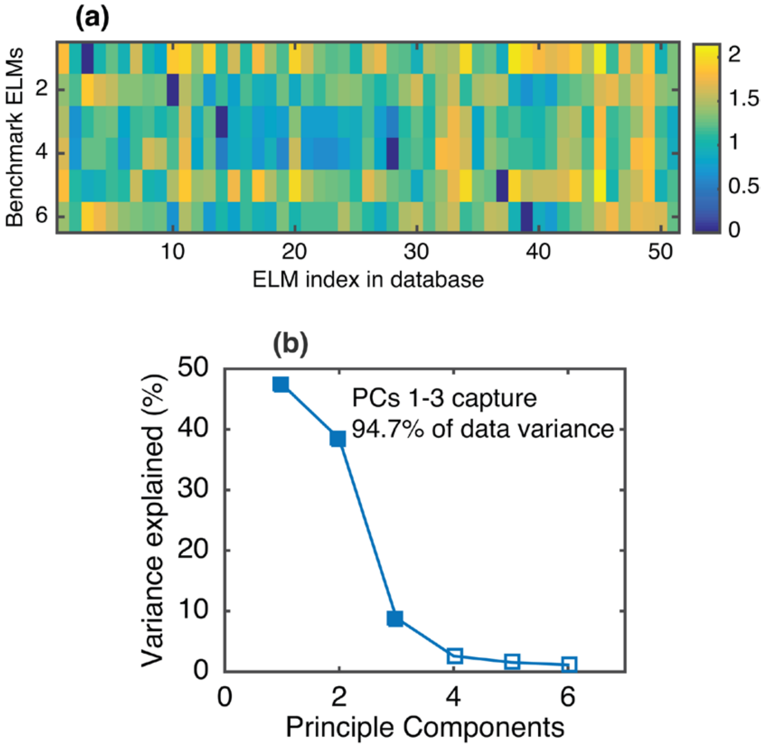

Figure 10. (a) Input data (geometric mean of four dissimilarity metrics) for *k*-means clustering with six benchmark ELM events. (b) Principal component analysis indicates the first three principal components capture nearly 95% of variation in the input data.

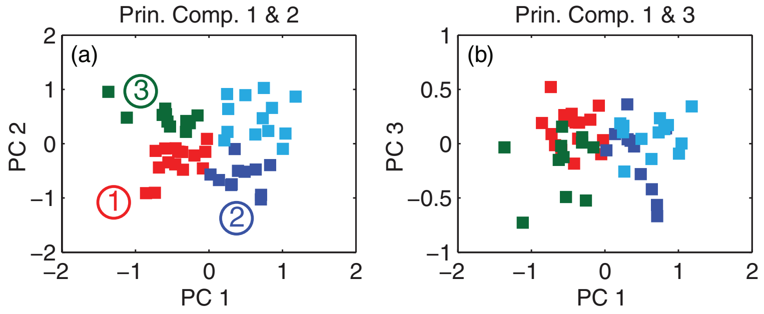

Figure 11. *k*-means cluster results with four clusters for six benchmark ELMs in figure 10. The clusters are plotted in principal component (PC) space to aid visualization: (a) clusters plotted in terms of PC 1 and PC 2, and (b) clusters plotted in terms of PC 1 and PC 3. The colors denote cluster membership.

Table 3. Optimal cluster number and S_J values with 6, 9, 11, and 14 of benchmark ELMs for *k*-means clustering.

No. of benchmark ELMs	Optimal no. of clusters	S_J
6	4	0.52
9	4	0.52
11	4	0.53
14	4	0.52

Note: For all benchmark scenarios, four clusters were optimal.

hierarchical results. As shown in figure 13, the clusters from *k*-means clustering largely map to clusters previously identified from hierarchical clustering. The cyan cluster from *k*-means analysis in figures 11 and 12 maps to a group of ELMs that are largely unlike all other ELMs from hierarchical analysis in figures 6, 7 and 9. In other words, *k*-means clustering captures the three clusters identified in hierarchical results plus a fourth cluster of ELMs (cyan) that defied grouping in the hierarchical results. Finally, figures 13(c)–(e) shows example ELMs from the identified clusters, and note that the cluster descriptions from figures 6(c)–(h) remain valid. We deliberately do not ascribe a number to the cyan cluster from *k*-means analysis because the cluster did not emerge as a low linkage cluster in hierarchical analysis. In other words, *k*-means and hierarchical analysis do not produce consistent results with regard to the cyan cluster in figure 13. In the next section, we explore parameter regimes for ELM clusters 1, 2, and 3 identified from hierarchical and *k*-means clustering.

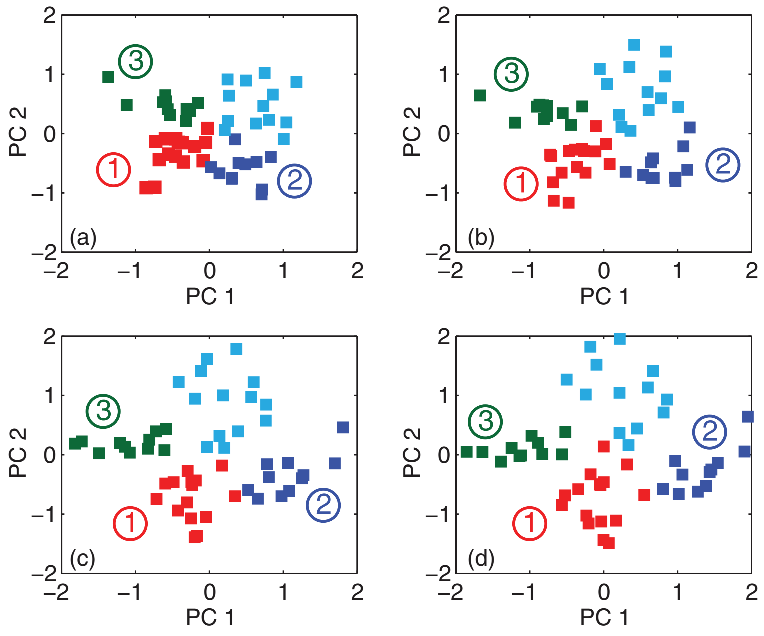

Figure 12. *k*-means clustering results with (a) 6, (b) 9, (c) 11, and (d) 14 benchmark ELMs. The clusters are nearly identical for all benchmark scenarios.

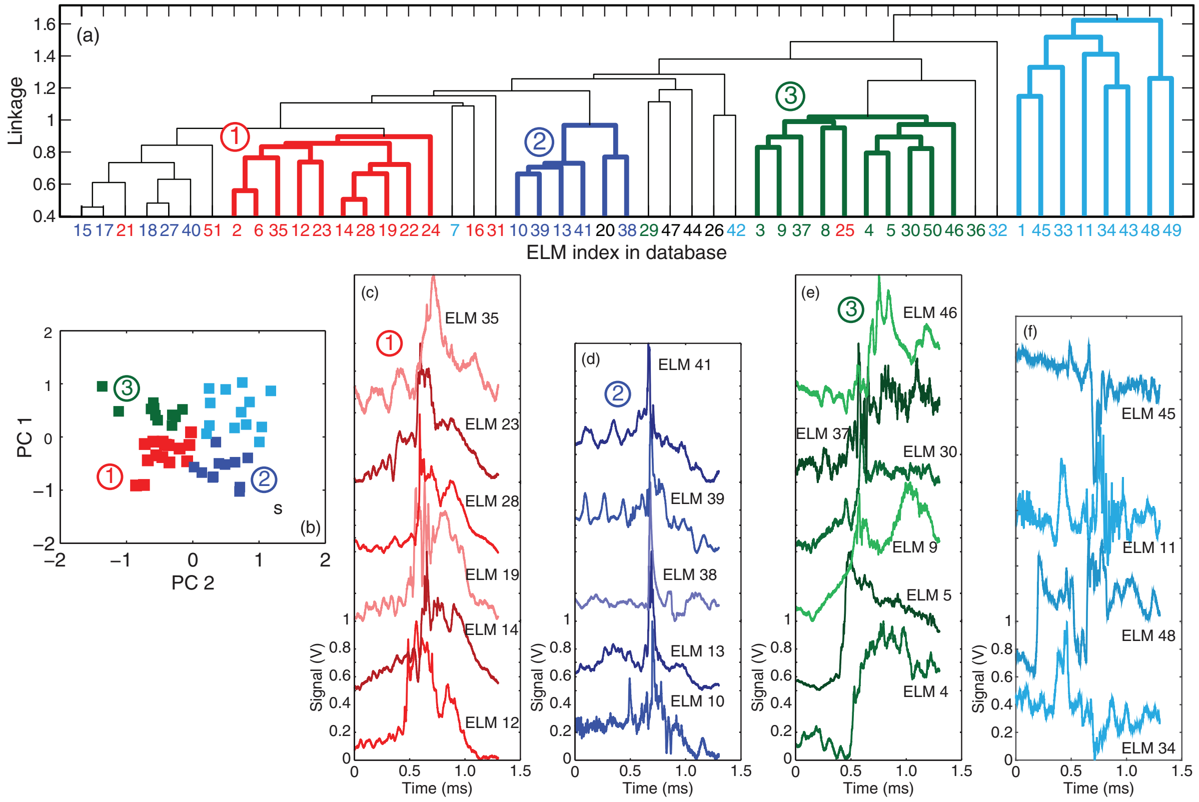

Figure 13. Comparison of (a) hierarchical and (b) *k*-means clustering results for the geometric mean of metrics with average linkage. The cyan cluster in (b) corresponds to the group of poorly matched ELMs in (a). The ELM number color in (a) corresponds to *k*-means cluster membership from (b). Example ELMs from (c) cluster 1, (d) 2, and (e) 3 and (f) the cyan group from *k*-means cluster results. Digital data for the ELM database can be found in [13].

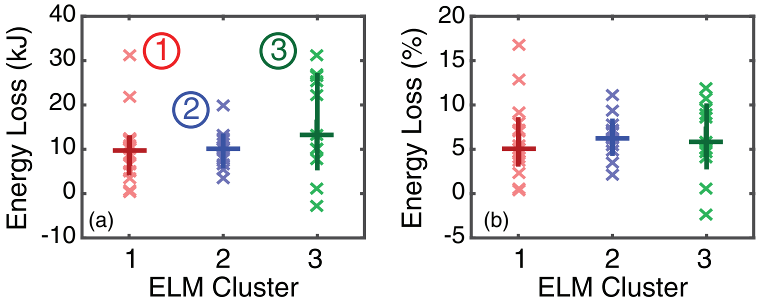

Figure 14. Stored energy losses for the ELM clusters in (a) kJ and (b) % loss. The small crosses (×) are individual ELMs, the solid bars are mean, 20th, and 80th percentile values, and colors and cluster numbers are consistent with figure 13. The 20th–80th percentile range captures typical parameter values. Digital data for the ELM database is available in [13].

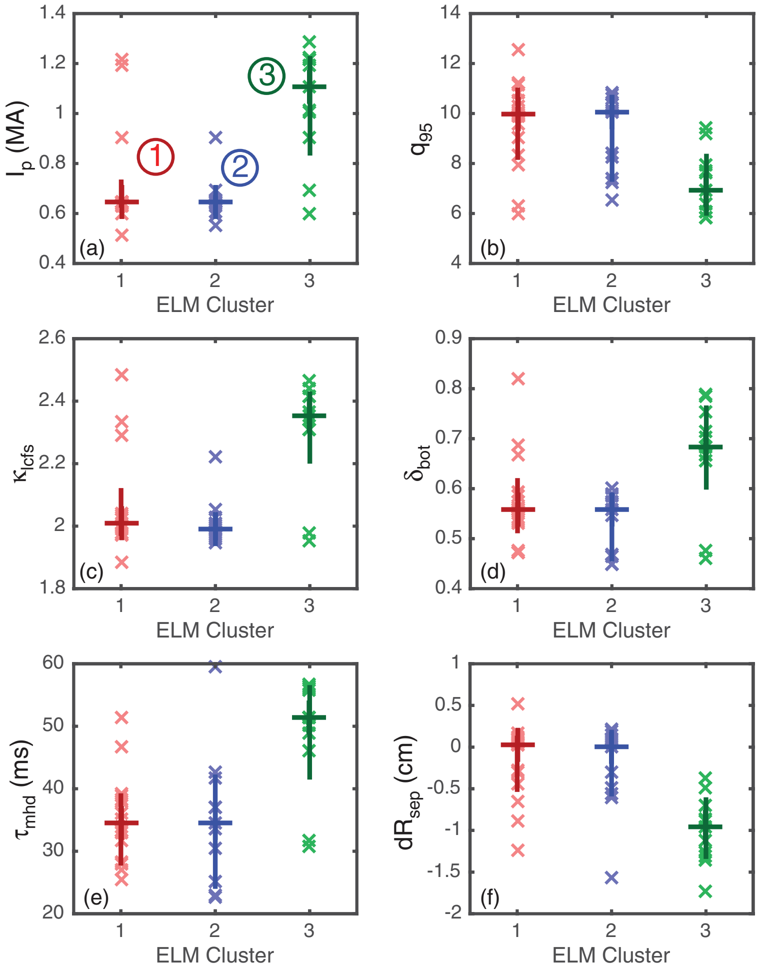

Figure 15. Solid bars denote mean, 20th, and 80th percentile values for equilibrium parameters for ELM clusters, and small crosses (×) denote values for individual ELMs: (a) plasma current, (b) q_{95} safety factor, (c) elongation at last-closed-flux-surface, (d) lower triangularity at last-closed-flux-surface, (e) stored energy, and (f) midplane separation of upper and lower X-points ($dR_{\mathrm{sep}} \lesssim -0.5$ cm is lower-single-null). The 20th–80th percentile range captures typical parameter values. Digital data for the ELM database is available in [13].

4. Parameter regimes for ELM clusters

Unsupervised clustering techniques identified three clusters of ELMs with similar evolution patterns in the previous section, and now we search for parameter regimes among ELM-relevant parameters that correlate with the identified ELM clusters. The observed evolution patterns reflect the nonlinear processes that impact ELM dynamics, and the corresponding parameter regimes can motivate theoretical or computational investigations of nonlinear ELM dynamics. Stored energy loss [18]

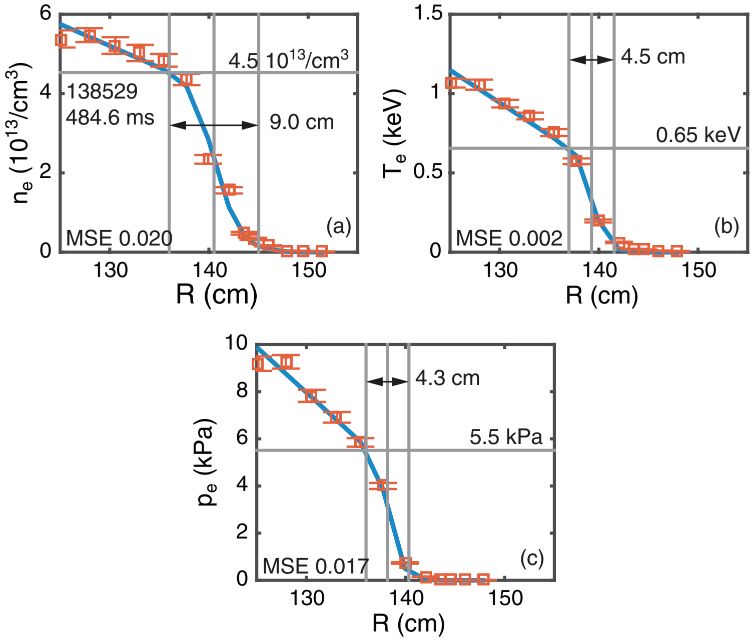

Figure 16. Example tanh fits to n_e, T_e, and p_e pedestal profiles from multi-point Thomson scattering measurements.

is a key metric for ELMs, but figure 14 indicates the identified ELM clusters exhibit similar stored energy losses. Some ELMs in figure 14 exhibit small or negative stored energy loss values due to ELM events with up to 15 ms of recovery time until the post-ELM magnetic reconstruction and Thomson scattering measurement (see figure 2(b)). The parameter ranges in figure 14 can be compared to table 1, but recall that table 1 covers approximate minimum and maximum values for the entire ELM dataset while figure 14 shows typical parameter ranges for clusters of ELMs with similar time-evolution characteristics. The ELM clusters do not appear to correlate with stored energy loss, but we continue the investigation with equilibrium parameters and pedestal profile characteristics.

Figure 15 shows equilibrium and magnetic reconstruction parameters for the ELM clusters [18]. Most notably, cluster 3 with prolonged elevated signals (see figures 6(c)–(h)) corresponds to higher I_p, and clusters 1 and 2, with shorter durations, correspond to lower I_p. The clustering with I_p is reminiscent of the fast and slow post-ELM pedestal temperature gradient recoveries observed in DIII-D [19]. Consistent with I_p patterns, cluster 3 also corresponds to lower safety factor q_{95}, lower magnetic shear, higher stored energy, and higher confinement time; clusters 1 and 2 correspond to the opposite parameter regimes. Large lower triangularity (δ_L) is a stabilizing factor for the linear peeling-ballooning mode [6], and we find cluster 3 preferentially occurs at higher δ_L values. In terms of geometry and magnetic balance, cluster 3 occurs preferentially at higher elongation (κ) and in lower single null configurations ($dR_{\text{sep}} \lesssim -0.5$ cm), and clusters 1 and 2 occur preferentially at lower elongation and double null configurations. The variations in ELM evolution could be due to geometry or magnetic topology variations, but regardless an accurate nonlinear model of ELM dynamics should capture variations in ELM evolution due to any factor including geometry or topology. Note, however, that the ELM database lacks observations in the upper single null configuration.

The pedestal width, height, and gradient are key quantities that impact pedestal stability, and the EPED model predicts the height and width of the pressure pedestal from constraints established by the peeling-ballooning mode and the kinetic ballooning mode (KBM) turbulence [20, 21]. Figure 16 shows pedestal heights and widths from tanh fits [22] to profiles for electron density, temperature, and pressure from multi-point Thomson scattering measurements [23]. The pedestal density gradient, for instance, is $\nabla n_e^{\text{ped}} \equiv n_e^{\text{ped}}/\Delta R_{n,\text{ped}}$. Figure 17 shows modest overlap in pedestal parameters for the ELM clusters, but we see that cluster 3 generally exhibits lower n_e^{ped}, higher Δn_e^{ped}, and smaller ∇n_e^{ped} values. In strongly shaped plasmas, higher density shifts the dominant peeling-ballooning mode to higher-n ballooning modes [6]. The dominance of clusters 1 and 2 at higher pedestal density values could be associated with a shift to higher-n ballooning modes. Recent results from JET indicate the post-ELM pedestal collapse time is longer in low T_e^{ped}, high collisionality regimes with the ITER-like wall [24]. The clusters in figure 17 exhibit similar ranges for T_e^{ped} and ν_{ei}, so the cluster results do not indicate T_e^{ped} and ν_{ei} are critical parameters for ELM evolution dynamics. Also, cluster 3 in figure 17(g) exhibits a larger ratio of temperature gradient to density gradient ($\nabla T_e^{\text{ped}}/\nabla n_e^{\text{ped}}$). The parameter patterns illustrated in figures 15 and 17 are consistent linear stability trends for the peeling-ballooning mode. For instance, lower pedestal density, higher triangularity, and higher plasma current are stabilizing for the peeling-ballooning mode [1]. In figures 15 and 17, the same parameter limits correspond to cluster 3, and the opposite parameter limits correspond to

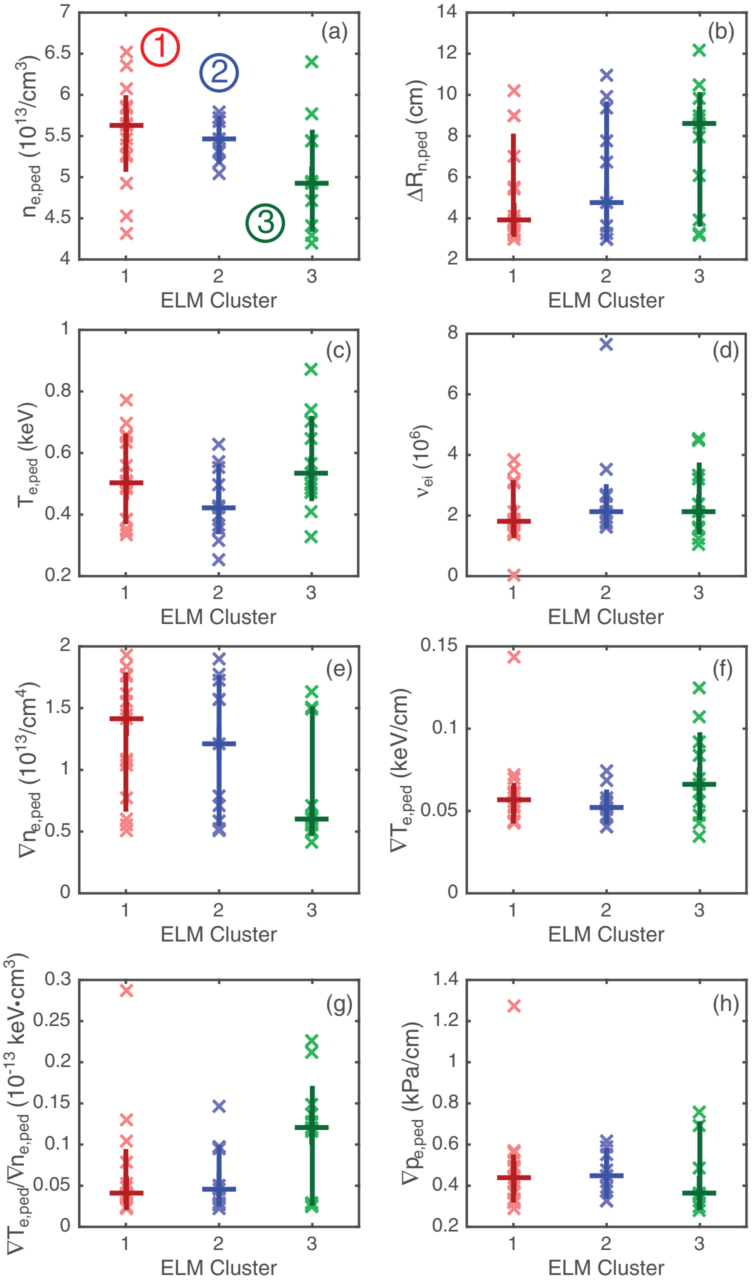

Figure 17. Solid bars denote mean, 20th, and 80th percentile values for pedestal parameters for ELM clusters, and small crosses (×) denote values for individual ELMs: (a) electron density pedestal height, (b) electron density pedestal width, (c) electron temperature pedestal height, (d) electron–ion collisionality, (e) density pedestal gradient ($\nabla n_e^{\rm ped} \equiv n_e^{\rm ped}/\Delta R_{n,\rm ped}$), (f) temperature pedestal gradient, (g) ratio of temperature to density pedestal gradients, and (h) pressure pedestal gradient. The 20th–80th percentile range captures typical parameter values. Digital data for the ELM database is available in [13].

clusters 1 and 2. Finally, parameters listed in table 1 but not shown in figures 15 and 17 exhibited similar ranges for the identified clusters.

Finally, we note that the clusters 1 and 2 correspond to similar parameter regimes in figures 15 and 17. Likewise, some clustering results (e.g. figure 9(d)) did not clearly distinguish clusters 1 and 2. Therefore, clusters 1 and 2 are perhaps best considered as a single cluster despite initial clustering results that pointed to distinct clusters. Nonetheless, clustering results and parameter regimes unambiguously point to at least two clusters of ELMs with distinct evolution patterns: (1) cluster 3 with longer ELM event duration and corresponding to higher plasma current, higher triangularity, lower-single-null configuration, lower $n_e^{\rm ped}$, and higher $\nabla T_e^{\rm ped}$, and (2) clusters 1/2 with shorter ELM event duration and corresponding to lower plasma current, lower triangularity, balanced double-null configuration, higher n_e^{ped}, and lower $\nabla T_e^{\mathrm{ped}}$. Based on the observed evolution patterns and parameter regimes, we expect the identified parameters will influence the evolution patterns and nonlinear dynamics in nonlinear simulations of ELM events.

5. Discussion of machine learning applications in fusion science

The preceding analysis demonstrates that unsupervised machine learning techniques can identify patterns or structure in data generated at data-rich fusion facilities. Pattern identification, a key step in scientific discovery, is typically accomplished with visual inspection of data, but visual inspection is not scalable to large or high-dimensional datasets. Unsupervised machine learning algorithms, however, can identify patterns or structure in large, complex datasets with computational speed and scalability. In addition, large, diverse datasets are less susceptible to selection bias and can generate results with broader relevance. Supervised machine learning techniques, on the other hand, quantify relationships among 'labeled' data, that is, data in which all relevant quantities or parameters are known. Supervised machine learning was recently applied to high-dimensional pedestal turbulence observations to quantify dozens of scaling relationships between turbulence quantities and plasma parameters [25, 26]. Coupling unsupervised and supervised machine learning techniques can, in principle, automate large portions of the scientific discovery workflow. For instance, unsupervised machine learning can identify patterns in data and a supervised learning algorithm can quantify relationships among the identified data groups. In fact, the entire analysis sequence in the previous sections for ELM pattern identification and parameter regimes could have been automated. Also, classification techniques from machine learning can scour data archives to identify new instances of an event or phenomenon prior to analysis with supervised learning algorithms. Machine learning techniques spanning pattern discovery, relationship quantification, and data classification present new opportunities to enhance the scientific productivity at data-rich experimental fusion facilities.

6. Summary

The linear peeling-ballooning stability boundary can capture ELM onset conditions, but ELM characteristics like intensity, filament dynamics, saturation mechanisms, and multi-mode interactions require nonlinear models and measurements with Alfvén-scale time resolution. Customary diagnostic tools, like multi-point Thomson scattering and filterscopes, cannot resolve dynamics on the Alfvén timescale. Also, heuristic ELM classification schemes (Type I, III, etc) based on extrinsic ELM properties, like secular edge emission and inter-ELM period, do not address the nonlinear dynamics and Alfvén-scale evolution of ELM events. In this paper, we investigated Alfvén-scale evolution patterns in ELM events captured by BES measurements on the National Spherical Torus Experiment, and digital data for this research activity are available in [13]. We implemented unsupervised machine learning algorithms that identified characteristic evolution patterns in a database of ELM events. Time-series similarity metrics (figures 5 and 8) quantified the similarity among ELM time-series data, and clustering algorithms (figures 9 and 11–13) identified two and possibly three clusters of ELMs with similar evolution characteristics. The ELM selection criteria for the database most likely admitted only Type I ELMs and excluded Type III and small, grassy ELMs. The identified ELM clusters triggered similar stored energy loss (figure 14), but the clusters occupied distinct parameter regimes for ELM-relevant parameters like plasma current, magnetic balance, triangularity, and pedestal height (figures 15 and 17). Notably, the pedestal electron pressure gradient is not an effective parameter for distinguishing the ELM groups, but the ELM groups are segregated in terms of electron density gradient and electron temperature gradient. Specifically, a cluster of ELM events (cluster 3 in figure 13) corresponds to longer ELM event duration, higher plasma current, higher triangularity, lower-single-null configuration, lower n_e^{ped}, and higher $\nabla T_e^{\mathrm{ped}}$, and another cluster of ELM events (clusters 1 and 2) correspond to shorter ELM event duration, lower plasma current, lower triangularity, balanced double-null configuration, higher n_e^{ped}, and lower $\nabla T_e^{\mathrm{ped}}$. The parameter regimes for the identified clusters connect to linear stability trends for the peeling-ballooning mode. Specifically, lower pedestal density, higher triangularity, and higher plasma current are stabilizing for the peeling-ballooning mode. The distinct evolution patterns and parameter regimes point to genuine variations in the underlying nonlinear dynamics. Based on the observed evolution patterns and parameter regimes, we expect the identified parameters will influence the evolution patterns and nonlinear dynamics in ELM simulations.

The analysis present here can be extended in several directions in future work. For instance, the evolution patterns can be templates for classification algorithms that automatically identify ELM instances in the data archive or a real-time data stream. Automated classification algorithms could populate an ELM database larger than anything possible with visual data inspection. Also, the algorithms and techniques could be extended or modified for other fast ELM-relevant measurements, such as magnetic or temperature fluctuations (though electron cyclotron emission measurements of temperature were not feasible in the low-field NSTX device). Extending hierarchical clustering to multiple fields (e.g. density and magnetic fluctuations) would require exploration of algorithms that combine multiple dissimilarity metrics into a single metric, but extending *k*-means clustering to multiple fields would be straightforward. Finally, the algorithms and techniques could be extended to other events such as Alfvén avalanches or disruptions. In summary, the analysis and results presented here demonstrate an application of unsupervised machine learning at a data-rich fusion facility, and a previous effort demonstrated a applications of supervised machine learning [25, 26]. Several scientific fields leverage machine learning techniques to automate discovery tasks in datasets too large or complex for comprehensive visual inspection. Machine learning techniques covering pattern identification, data classification, and relationship quantification offer new strategies for scalable and automated scientific discovery at data-rich fusion facilities.

Acknowledgments

The author DRS acknowledges helpful discussions with R Maingi. This material is based upon work supported by the US Department of Energy, Office of Science, Office of Fusion Energy Sciences under Award Numbers DE-FG02-89ER53296, DE-SC0001288, and DE-AC02-09CH11466. This research used resources of the National Spherical Torus Experiment-Upgrade, which is a DOE Office of Science User Facility. The digital data for this paper can be found in [13].

References

[1] Snyder P B, Wilson H R, Osborne T H and Leonard A W 2004 *Plasma Phys. Control. Fusion* **46** A131
[2] Snyder P B, Wilson H R and Xu X Q 2005 *Phys. Plasmas* **12** 056115
[3] Xu X Q, Dudson B, Snyder P B, Umansky M V and Wilson H 2010 *Phys. Rev. Let.* **105** 175005
[4] Xu X Q, Dudson B D, Snyder P B, Umansky M V, Wilson H R and Casper T 2011 *Nucl. Fusion* **51** 103040
[5] Holzl M *et al* 2012 *Phys. Plasmas* **19** 082505
[6] Snyder P B *et al* 2007 *Nucl. Fusion* **47** 961
[7] Burrell K H *et al* 2009 *Nucl. Fusion* **49** 085024
[8] Leonard A W 2014 *Phys. Plasmas* **21** 090501
[9] Smith D R, Feder H, Feder R, Fonck R J, Labik G, McKee G R, Schoenbeck N, Stratton B C, Uzun-Kaymak I and Winz G 2010 *Rev. Sci. Instrum.* **81** 10D717
[10] Smith D R, Fonck R J, McKee G R and Thompson D S 2012 *Rev. Sci. Instrum.* **83** 10D502
[11] Ono M *et al* 2000 *Nucl. Fusion* **40** 557
[12] Kaye S M *et al* 2001 *Phys. Plasmas* **8** 1977
[13] http://arks.princeton.edu/ark:/88435/dsp01h415pc93f
[14] Xu R and Wunsch D 2005 *IEEE Trans. Neural Netw.* **16** 645
[15] Eisen M B, Spellman P T, Brown P O and Botstein D 1998 *Proc. Natl Acad. Sci. USA* **95** 14863
[16] Liao T W 2005 *Pattern Recognit.* **38** 1857
[17] Mallat S G 1989 *IEEE Trans. Pattern Anal. Mach. Intel.* **11** 674
[18] Sabbagh S *et al* 2001 *Nucl. Fusion* **41** 1601
[19] Diallo A, Groebner R J, Rhodes T L, Battaglia D J, Smith D R, Osborne T H, Canik J M, Guttenfelder W and Snyder P B 2015 *Phys. Plasmas* **22** 056111
[20] Snyder P B *et al* 2009 *Phys. Plasmas* **16** 056118
[21] Snyder P B *et al* 2011 *Nucl. Fusion* **51** 103016
[22] Groebner R J *et al* 2001 *Nucl. Fusion* **41** 1789
[23] LeBlanc B P, Diallo A, Labik G and Stevens D R 2012 *Rev. Sci. Instrum.* **83** 10D527
[24] Frassinetti L *et al* 2015 *Nucl. Fusion* **55** 023007
[25] Smith D R *et al* 2013 *Phys. Plasmas* **20** 055903
[26] Smith D R *et al* 2013 *Nucl. Fusion* **53** 113029